ВВЕДЕНИЕ В БИОИНФОРМАТИКУ

Лекция №8

Скрытые марковские модели: Алгоритмы и применение

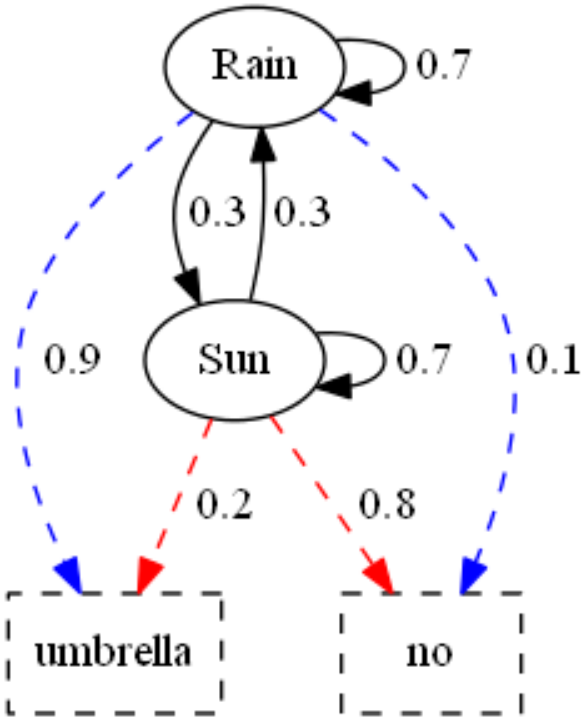
Новоселецкий Валерий Николаевич
к.ф.-м.н., доц. каф. биоинженерии
valery.novoseletsky@yandex.ru

Сайт курса http://intbio.org/bioinf2020-21

Погода

За день погода может поменяться с вероятностью 0,3, и если на улице идет дождь, то некий человек приносит зонтик с вероятностью 0,9, а если солнечно — то с вероятностью 0,2.

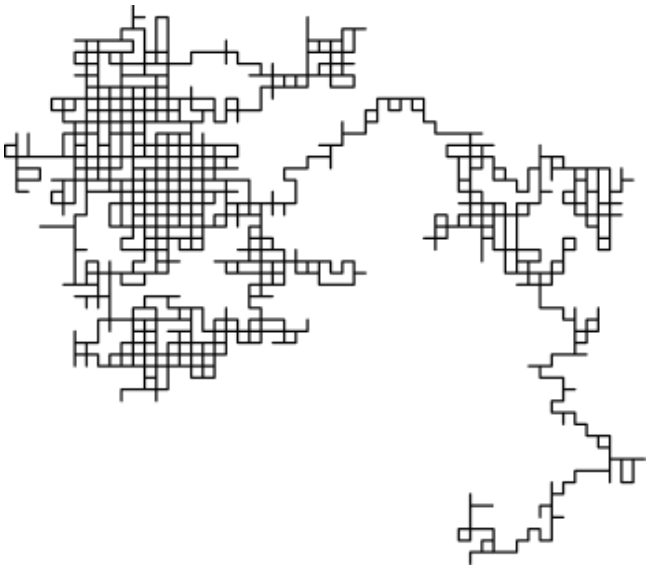
За рабочую неделю вы заметили, что он не принес зонтик лишь в среду. С какой вероятностью во вторник шел дождь?

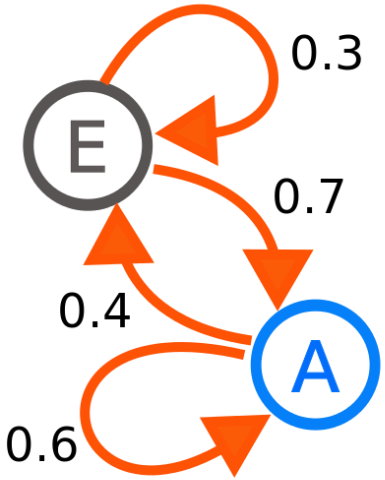


Марковский процесс

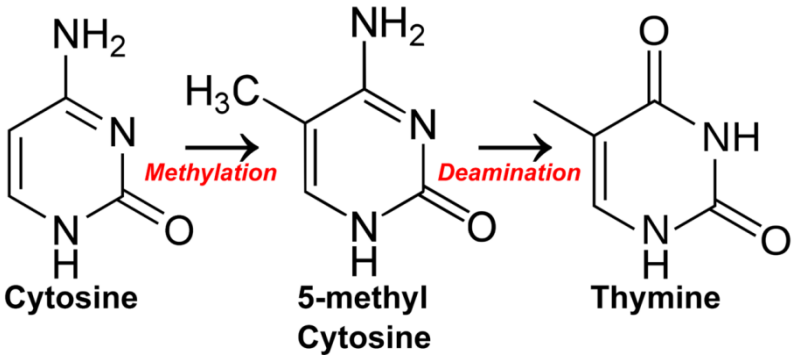
Марковский процесс — случайный процесс, эволюция которого после любого заданного значения временно́го параметра *t* не зависит от эволюции, предшествовавшей *t*, при условии, что значение процесса в этот момент фиксировано («будущее» процесса не зависит от «прошлого» при известном «настоящем») (бросание/перекатывание игрального кубика, случайное блуждание,…).

А.А. Марков (ст.)
(1856 -1922)





CpG-островки

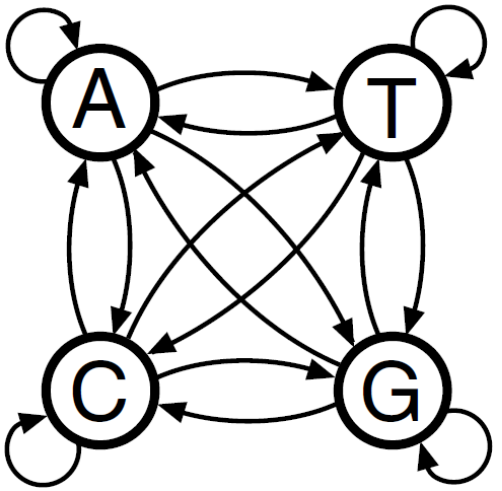


CpG

GpC

CpG-островки

Задача 1: Дан небольшой фрагмент последовательности.
Принадлежит ли он CpG-островку?



Цепь Маркова для ДНК

$a_{st} = \mathrm{P}(\mathrm{x_i} = \mathrm{t} \mid \mathrm{x_{i-1}} = \mathrm{s})$ - вероятности перехода

Вероятность появления последовательности x

$$P(x) \equiv P(x_L, x_{L-1}, \ldots, x_1) = P(x_L | x_{L-1}, \ldots, x_1) P(x_{L-1} | x_{L-2}, \ldots, x_1) \ldots P(x_1)$$

Или, принимая во внимание свойство цепи Маркова,

$$P(x) = P(x_L | x_{L-1}) P(x_{L-1} | x_{L-2}) \ldots P(x_2 | x_1) P(x_1) = P(x_1) \prod_{i=2}^{L} a_{x_{i-1} x_i}$$

CpG-островки

Рассмотрим набор фрагментов последовательности ДНК человека (60 000 нуклеотидов).

Построим две модели: для фрагментов, являющихся CpG-островками («+»-модель), и для фрагментов, не являющихся («-»-модель) (фрагменты классифицированы заранее).

Анализируя фрагменты, рассчитаем вероятности переходов $a_{st} = \frac{c_{st}}{\sum_u c_{su}}$

+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

−	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

Переходы C->G в обеих моделях ожидаемо наблюдаются реже, чем переходы G->C, причём в «-»-модели эффект выражен сильнее.

CpG-островки

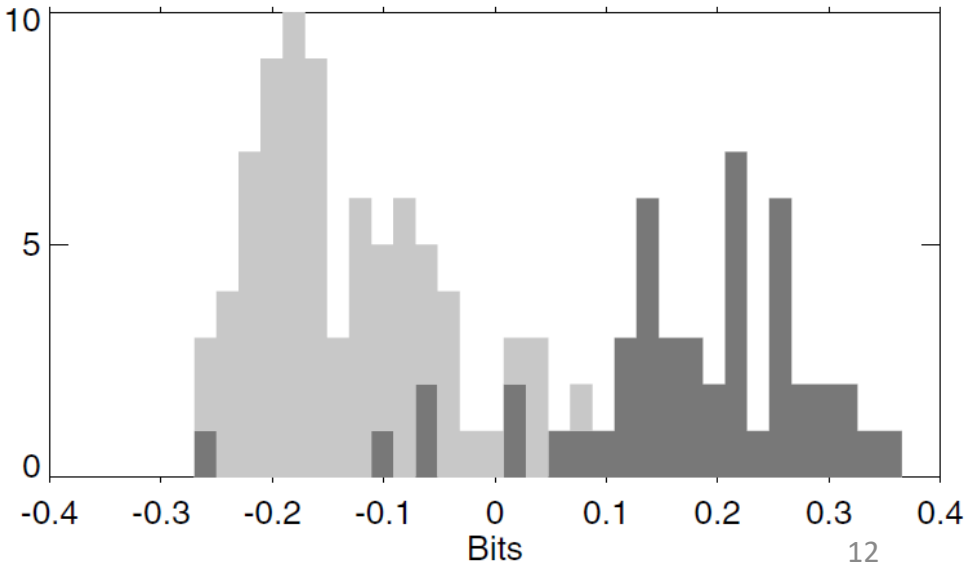
За оценку фрагмента примем логарифм отношения правдоподобия его появления:

$$S(x) = \log \frac{P(x|\text{model}+)}{P(x|\text{model}-)} = \sum_{i=1}^{L} \log \frac{a^{+}_{x_{i-1}x_i}}{a^{-}_{x_{i-1}x_i}} = \sum_{i=1}^{L} \beta_{x_{i-1}x_i}$$

В случае использования двоичного логарифма для β имеем таблицу в битах:

β	A	C	G	T
A	−0.740	0.419	0.580	−0.803
C	−0.913	0.302	1.812	−0.685
G	−0.624	0.461	0.331	−0.730
T	−1.169	0.573	0.393	−0.679

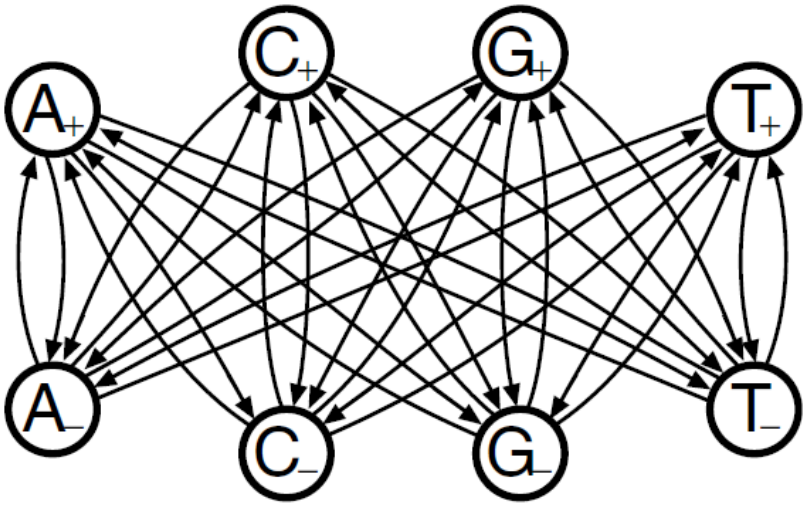
Теперь для <нормированных на длину> оценок фрагментов получаем распределение: оценки фрагментов с CpG-островками в среднем заметно больше, чем фрагментов без них.



CpG-островки

Задача 2: Дан большой фрагмент последовательности.
Как найти в нём CpG-островки?

Цепь Маркова для ДНК теперь содержит состояния «+» и «-», характеризуемые небольшими вероятностями переходов:

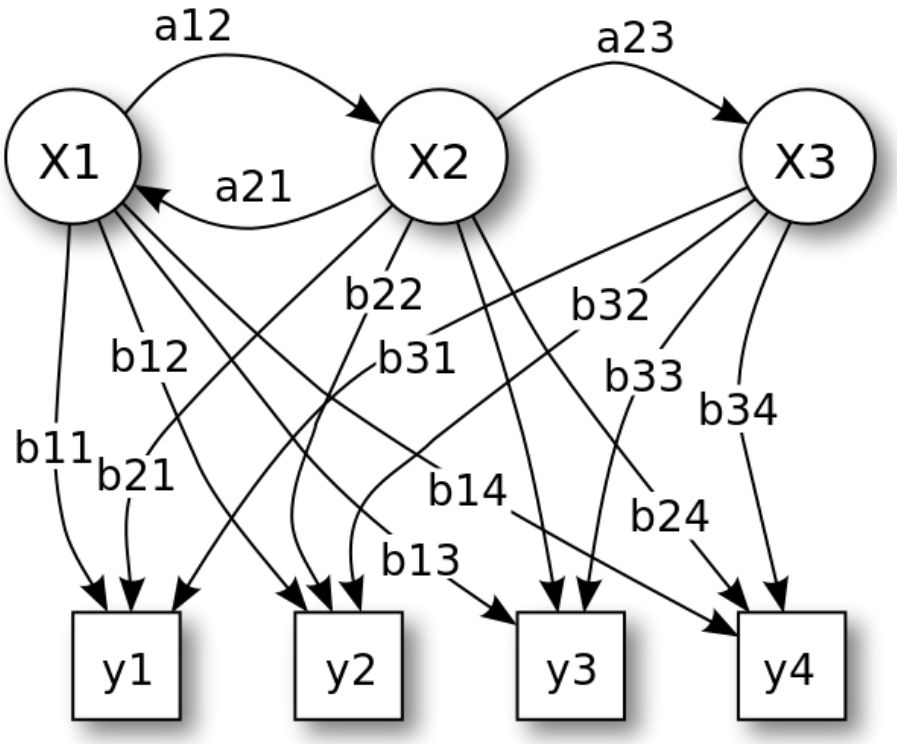


(Переходы внутри каждой из групп для наглядности не показаны)

При рассмотрении символа C в последовательности невозможно определить, был ли он порождён моделью «+» или «-».

Скрытые марковские модели

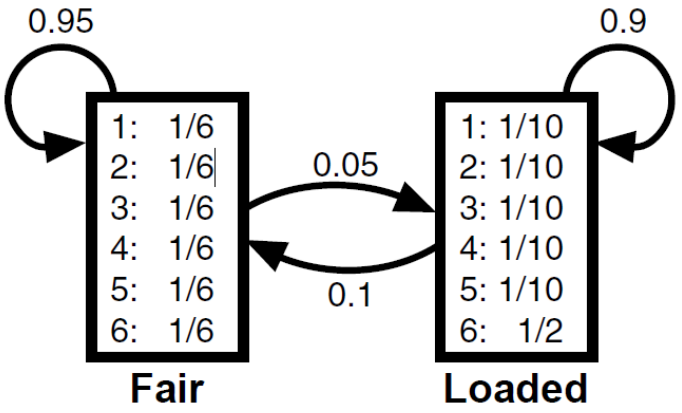
Скрытая марковская модель (СММ, hidden Markov model, HMM) — статистическая модель, имитирующая работу процесса, похожего на марковский процесс с неизвестными параметрами. Задачей ставится разгадывание неизвестных параметров на основе наблюдаемых.



Казино

Вы играете с игроком от казино:

1. Делаются ставки;
2. Вы бросаете кубик;
3. Соперник бросает кубик;
4. Обе ставки забирает тот, у кого выпало больше.



Казино

При ряде бросков у Вашего соперника выпала следующая последовательность:

12156216241461461361366616646616366163661636165

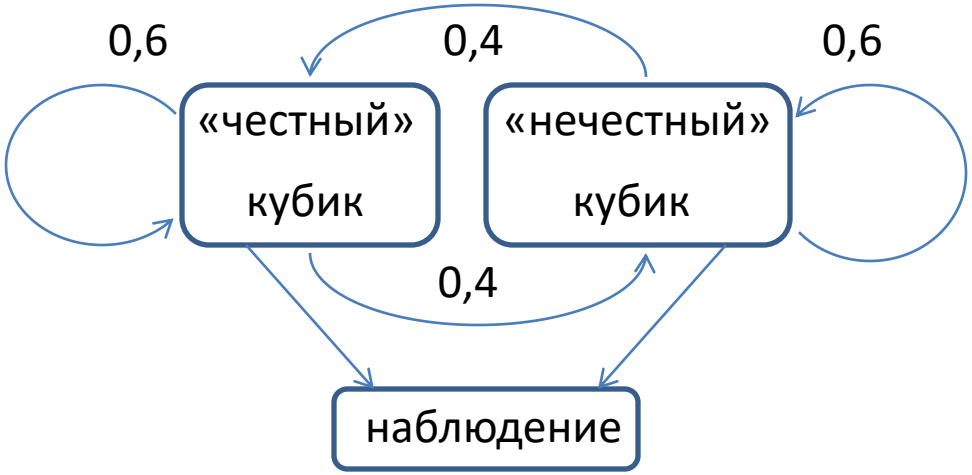
Какие вопросы могут у Вас возникнуть?

1. Насколько вероятно выпадение такой последовательности в рамках известной модели казино?
   В терминах СММ это **задача ОЦЕНКИ**
2. Считая модель верной, какие фрагменты последовательности могли быть сгенерированы «честным» кубиком, а какие «нечестным»?
   В терминах СММ это **задача ДЕШИФРОВКИ**
3. Насколько смещён центр тяжести в «нечестном» кубике? Насколько идеален «честный» кубик? Как часто соперник меняет кубики?
   Это **задача** определения параметров или **ОБУЧЕНИЯ** и она самая сложная.

Казино. Задача оценки

Рассмотрим скрытую марковскую модель смены кубиков:

P(1|Ч) = 1/6
P(2|Ч) = 1/6
P(3|Ч) = 1/6
P(4|Ч) = 1/6
P(5|Ч) = 1/6
P(6|Ч) = 1/6



P(1|Н) = 1/10
P(2|Н) = 1/10
P(3|Н) = 1/10
P(4|Н) = 1/10
P(5|Н) = 1/10
P(6|Н) = 1/2

Какова вероятность того, что последовательность результатов 1, 2, 1, 5, 6, 2, 1, 6, 2, 4 получена при бросании «честного» кубика?

Решение:

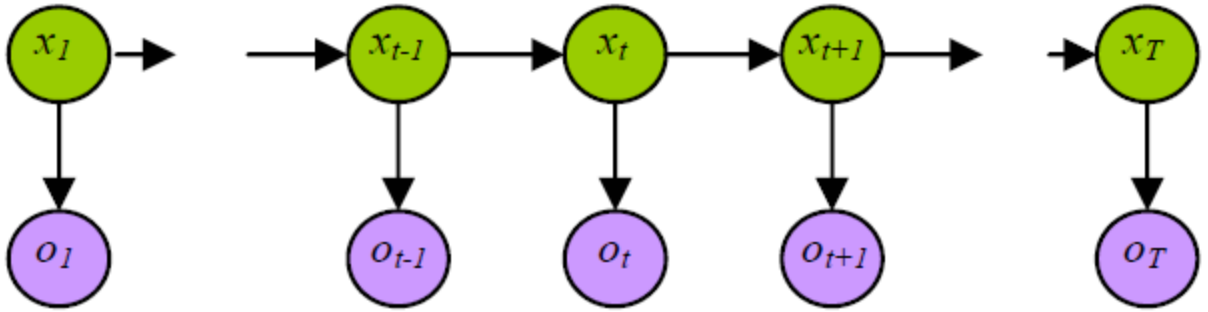
½* P(1|Ч) * P(Ч|Ч) * P(2|Ч) * P(Ч|Ч) … P(4|Ч) = ½* $(1/6)^{10}$ * $(0,6)^{9}$ = 8,3e-11

Какова вероятность того, что эта последовательность получена при бросании «нечестного» кубика?

½* P(1|Н) * P(Н|Н) * P(2|Н) * P(Н|Н) … P(4|Н) = ½ * $(1/10)^{8}$ * $(1/2)^{2}$ * $(0,6)^{9}$ = 1,3e-11

Т.о. P(ЧЧЧЧЧЧ) = P(НННННН)*6,6, но легко подобрать и обратный пример (как?)

Формализация

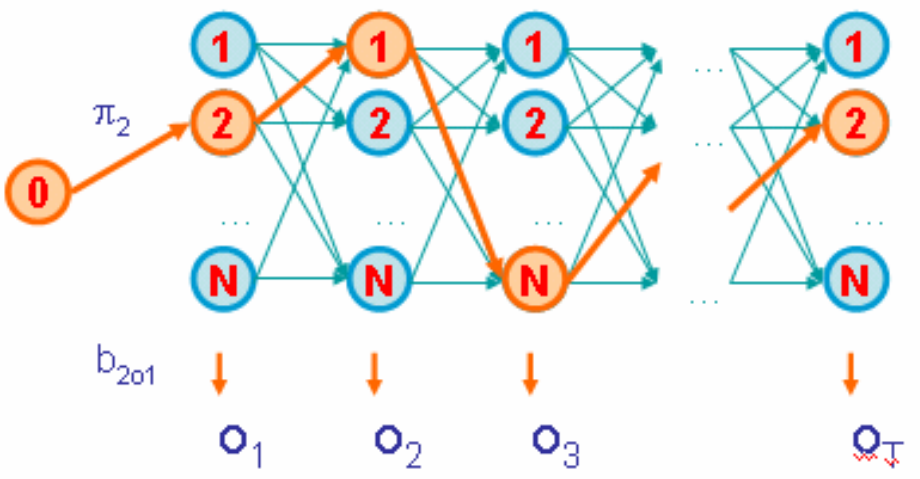


СММ μ может быть задана с помощью 3 матриц {П, А, В}:

П = {π_i} – вероятности начальных состояний

А = {a_{ij}} – вероятности переходов, т.е. Pr(x_j|x_i) – вероятность перехода из состояния x_i в состояние x_j.

В = {b_{ik}} – вероятности наблюдений, т.е. Pr(o_k|x_i) – вероятность порождения наблюдения o_k состоянием x_i.



Задача оценки

Имея СММ μ = {П, А ,В} и последовательность наблюдений О = {О1, …, Оτ}, вычислить Pr[O|μ] – вероятность порождения последовательности наблюдений данной моделью.

Решение:

Пусть Х = {Х1, …, Хτ} – последовательность состояний.

Легко видеть, что

$$\Pr(X|\mu) = \pi_{x_1} a_{x_1 x_2} a_{x_2 x_3} \dots a_{x_{n-1} x_n} = \pi_{x_1} \prod_{t=1}^{T-1} a_{x_t x_{t+1}} \quad (1)$$

Аналогично

$$\Pr(O|X,\mu) = b_{x_1 o_1} b_{x_2 o_2} \dots b_{x_T o_T} = b_{x_1 o_1} \prod_{t=1}^{T-1} b_{x_{t+1} o_{t+1}} \quad (2)$$

Переходя к рассмотрению всех возможных последовательностей состояний, получаем

$$\Pr(O|\mu) = \sum_X \Pr(O, X|\mu) = \sum_X \Pr(O|X,\mu) \Pr(X|\mu) \quad (3)$$

Подстановка (1) и (2) в (3) даёт

$$= \sum_X \pi_{x_1} \prod_{t=1}^{T-1} a_{x_t x_{t+1}} \bullet b_{x_1 o_1} \prod_{t=1}^{T-1} b_{x_{t+1} o_{t+1}} = \sum_X \pi_{x_1} b_{x_1 o_1} \prod_{t=1}^{T-1} a_{x_t x_{t+1}} b_{x_{t+1} o_{t+1}}$$

Все параметры известны, но количество вычислений составляет O(2TN^T)

= > **нужен более эффективный алгоритм!**

Задача дешифровки

Имея СММ μ = {П, А ,В} и последовательность наблюдений О = {О1, …, Оτ}, найти наиболее вероятную последовательность состояний Х = {Х1, …, Хτ}, порождающую эту последовательность наблюдений.

$$\hat{X} = \arg\max_{X} \Pr(X \mid O, \mu)$$

Решение:

Сперва изменим формулировку. Поскольку

$$\max \Pr(X \mid O, \mu) = \frac{\max \Pr(X, O \mid \mu)}{\Pr(O \mid \mu)}$$

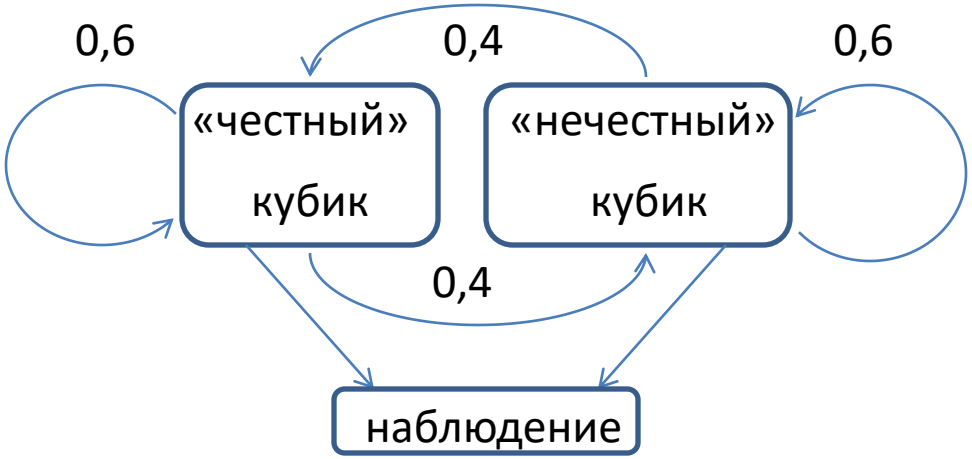
а знаменатель не зависит от Х, то

$$\hat{X} = \arg\max_{X} \Pr(X \mid O, \mu) = \arg\max_{X} \Pr(X, O \mid \mu)$$

Алгоритм Витерби (1967)

В терминах модели смены кубиков:

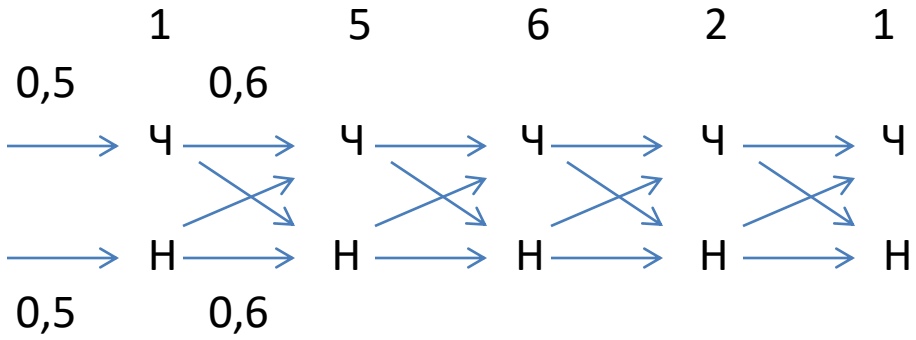
P(1|Ч) = 1/6
P(2|Ч) = 1/6
P(3|Ч) = 1/6
P(4|Ч) = 1/6
P(5|Ч) = 1/6
P(6|Ч) = 1/6



P(1|Н) = 1/10
P(2|Н) = 1/10
P(3|Н) = 1/10
P(4|Н) = 1/10
P(5|Н) = 1/10
P(6|Н) = 1/2

Найти наиболее вероятную последовательность состояний кубика, порождающую последовательность результатов 1, 5, 6, 2, 1.

Решение:

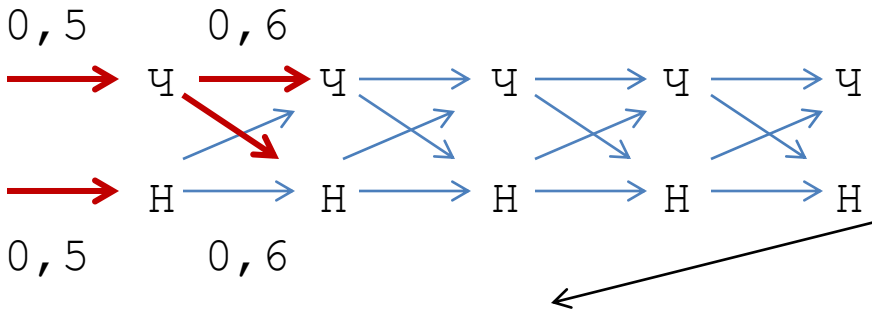


Алгоритм Витерби (1967)

Найти наиболее вероятную последовательность состояний кубика, порождающую последовательность результатов 1, 5, 6, 2, 1.

Решение (продолжение):

```
T:      1      2      3      4      5  (номера состояний)
O:      1      5      6      2      1  (наблюдения)
```



вероятности того, что данное наблюдение стало результатом броска того или иного кубика

T=1: P(Ч1) = ½*1/6 = 0,0833 P(Н1) = ½ * 1/10 = 0,05

T=2: **P(Ч1Ч) = ½*1/6*0,6 = 0,050 = > P(Ч1Ч5) = ½*1/6*0,6*1/6 = 0,0083**

P(Н1Ч) = ½*1/10*0,4 = 0,020

P(Ч1Н) = ½*1/6*0,4 = 0,033 = > P(Ч1Н5) = ½*1/10*0,4*1/10 = 0,0033

P(Н1Н) = ½*1/10*0,6 = 0,030

Алгоритм Витерби (1967)

Найти наиболее вероятную последовательность состояний кубика, порождающую последовательность результатов 1, 5, 6, 2, 1.

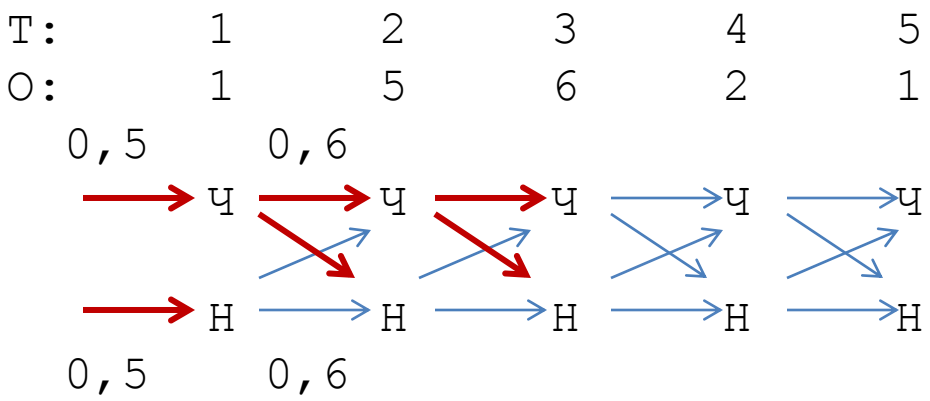
Решение (продолжение):

T=3: **P(Ч1Ч5Ч) = 0,0083*0,6 = 5,0e-3 => P(Ч1Ч5Ч6) = 5,0e-3 *1/6 = 8,3e-4**

P(Ч1Н5Ч) = 0,0033*0,4 = 1,3e-3

P(Ч1Ч5Н) = 0,0083*0,4 = 3,3e-3 => P(Ч1Ч5Н6) = 3,3e-3 *1/2 = 1,7e-3

P(Ч1Н5Н) = 0,0033*0,6 = 2,0e-3



Алгоритм Витерби (1967)

Найти наиболее вероятную последовательность состояний кубика, порождающую последовательность результатов 1, 5, 6, 2, 1.

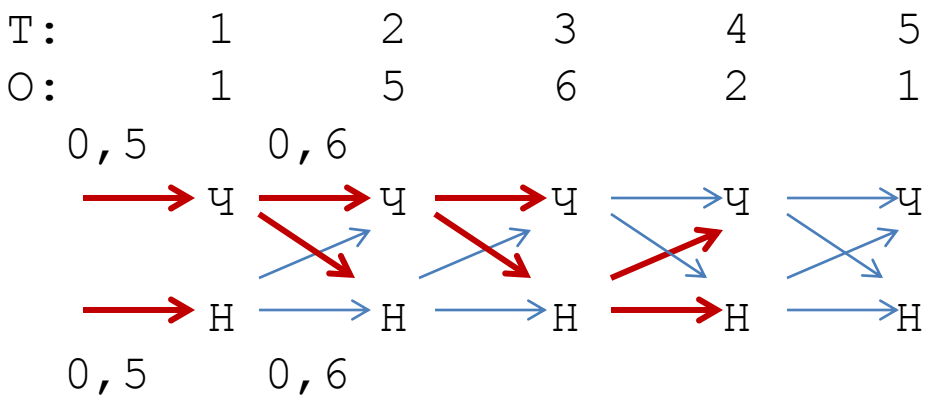
Решение (продолжение):

T=4: P(Ч1Ч5Ч6Ч) = 8,3e-4*0,6 = 5,0e-4

P(Ч1Ч5Н6Ч) = 1,7e-3*0,4 = 6,7e-4 => P(Ч1Ч5Н6Ч2) = 6,7e-4*1/6 = 1,1E-4

P(Ч1Ч5Ч6Н) = 8,3e-4*0,4 = 3,3e-4

P(Ч1Ч5Н6Н) = 1,7e-3*0,6 = 1,0e-3 => P(Ч1Ч5Н6Н2) = 1,0e-3*1/10=1,0E-4



Алгоритм Витерби (1967)

Найти наиболее вероятную последовательность состояний кубика, порождающую последовательность результатов 1, 5, 6, 2, 1.

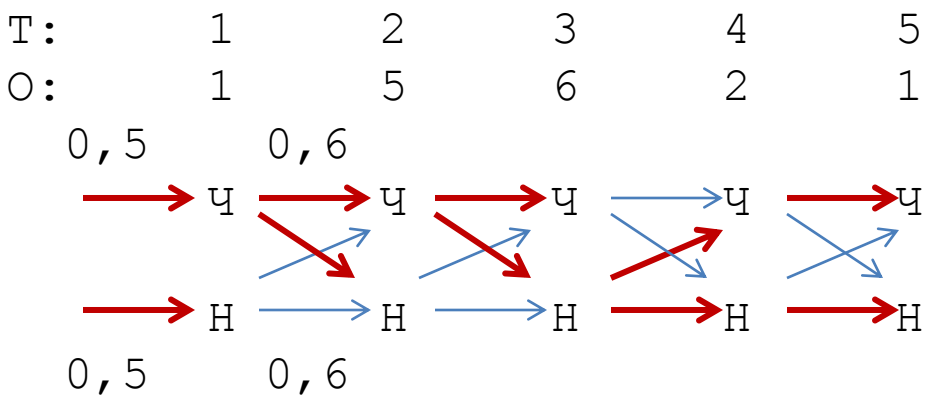
Решение (окончание):

T=5: **P(Ч1Ч5Н6Ч2Ч) = 1,1E-4*0,6 = 6,7e-5 => P(Ч1Ч5Н6Ч2Ч1) = 6,7e-5*1/6 = 1,1E-5**

P(Ч1Ч5Н6Н2Ч) = 1,0E-4*0,4 = 4,0e-5

P(Ч1Ч5Н6Ч2Н) = 1,1E-4*0,4 = 4,4e-5

P(Ч1Ч5Н6Н2Н) = 1,0E-4*0,6 = 6,0e-5 => P(Ч1Ч5Н6Н2Н1) = 6,0e-5*1/10 = 6,0e-6



Алгоритм Витерби (1967)

Найти наиболее вероятную последовательность состояний кубика, порождающую последовательность результатов 1, 5, 6, 2, 1.

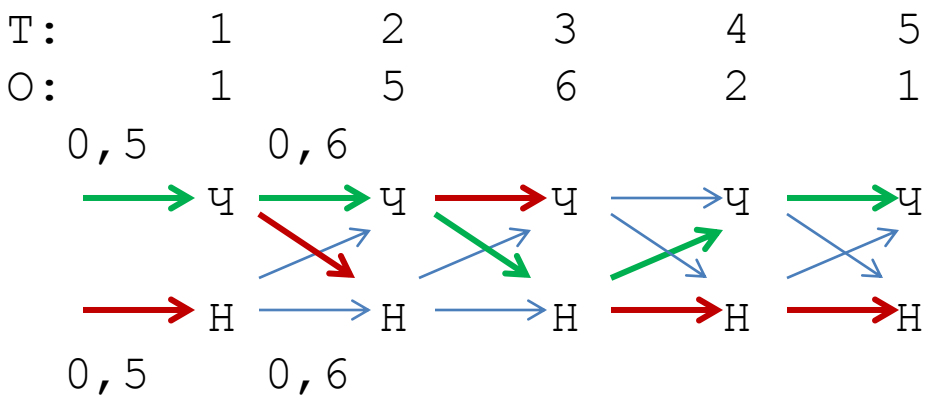
Решение (окончание):

T=5: **P(Ч1Ч5Н6Ч2Ч) = 1,1E-4*0,6 = 6,7e-5 => P(Ч1Ч5Н6Ч2Ч1) = 6,7e-5*1/6 = 1,1E-5**

P(Ч1Ч5Н6Н2Ч) = 1,0E-4*0,4 = 4,0e-5

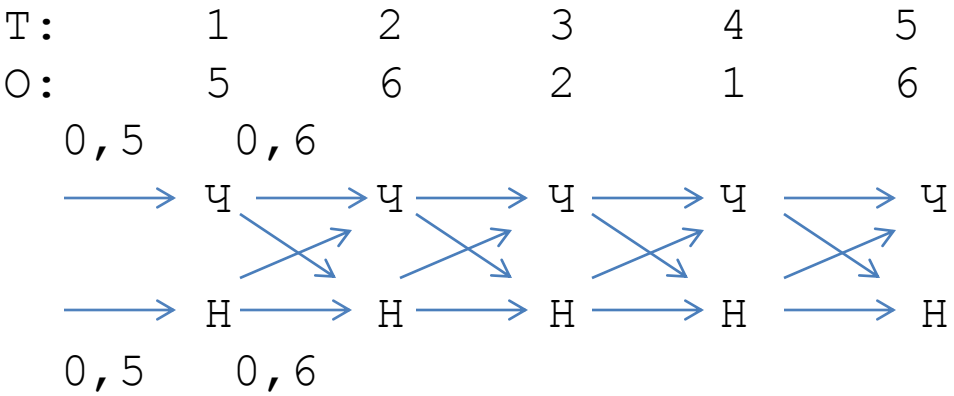
P(Ч1Ч5Н6Ч2Н) = 1,1E-4*0,4 = 4,4e-5

P(Ч1Ч5Н6Н2Н) = 1,0E-4*0,6 = 6,0e-5 => P(Ч1Ч5Н6Н2Н1) = 6,0e-5*1/10 = 6,0e-6



Алгоритм Витерби (1967)

Другая последовательность результатов 5, 6, 2, 1, 6.



T=1: P(Ч5) = ½*1/6 = 0,0833 P(Н5) = ½*1/10 = 0,05

Дальше – самостоятельно ☺

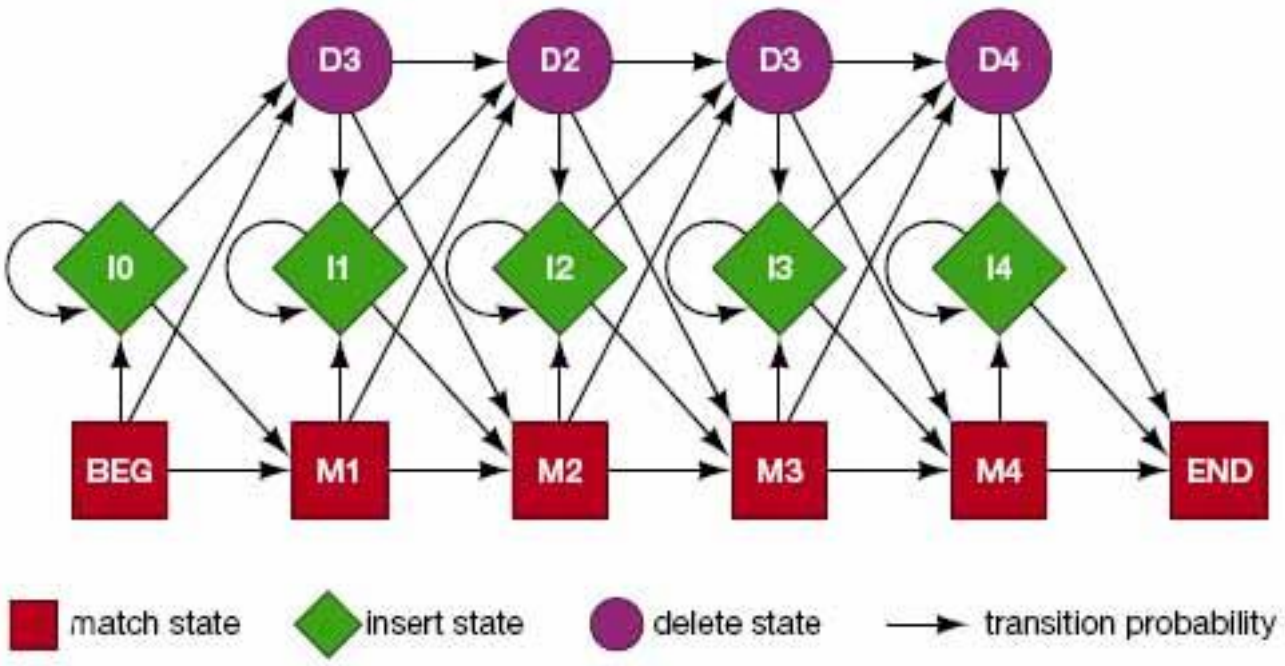
Построение множественного выравнивания

A. Sequence alignment

N	•	F	L	S
N	•	F	L	S
N	K	Y	L	T
Q	•	W	–	T

RED POSITION REPRESENTS ALIGNMENT IN COLUMN
GREEN POSITION REPRESENTS INSERT IN COLUMN
PURPLE POSITION REPRESENTS DELETE IN COLUMN

B. Hidden Markov model for sequence alignment



Построение множественного выравнивания

Алгоритм:

- **Обучение.** Имея ряд невыровненных последовательностей, можно выровнять их и подогнать вероятности переходов и порождения остатков, чтобы определить модель, описывающую данный набор последовательностей.
- **Поиск гомологов.** Имея модель и исследуемую последовательность, можно посчитать вероятность того, то модель могла бы сгенерировать эту последовательность. Если вероятность достаточно высока, то рассматриваемая последовательность принадлежит тому же семейству, что и обучающие.

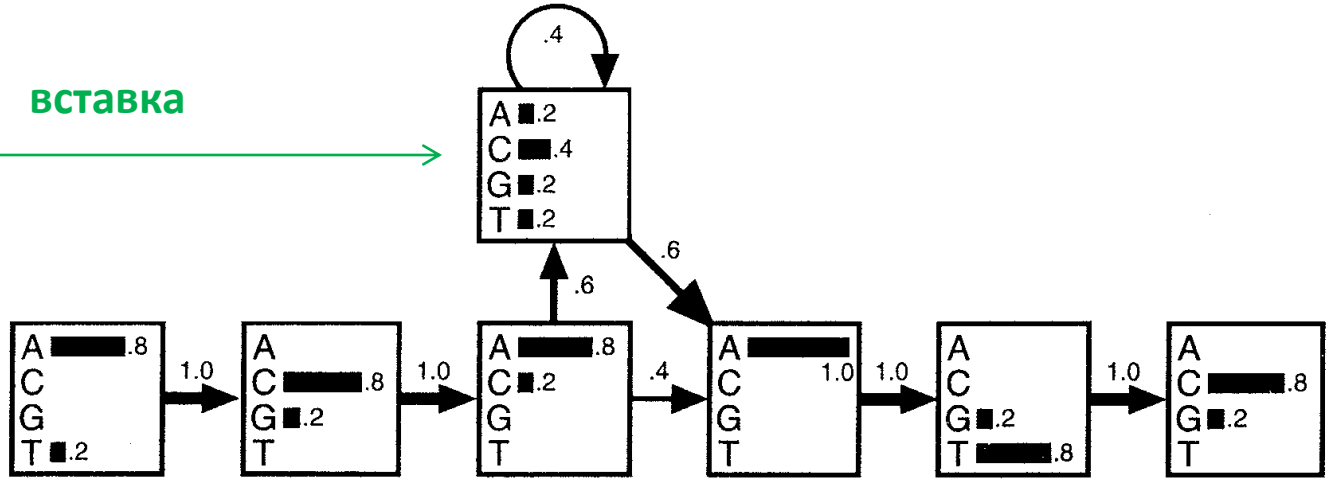
```
ACA---ATC
TCAACTATC
ACAC--AGC
AGA---ATC
ACCG--ATG
```

Построим?

Построение множественного выравнивания

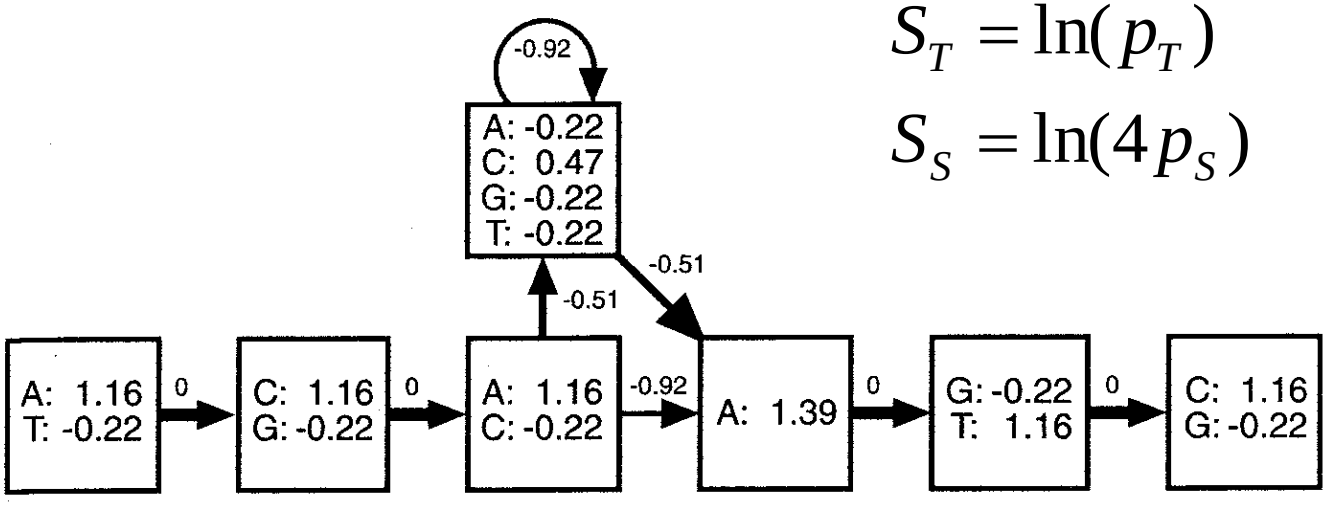
```
ACA---ATC
TCAACTATC
ACAC--AGC
AGA---ATC
ACCG--ATG
```

вставка



$$S_T = \ln(p_T)$$

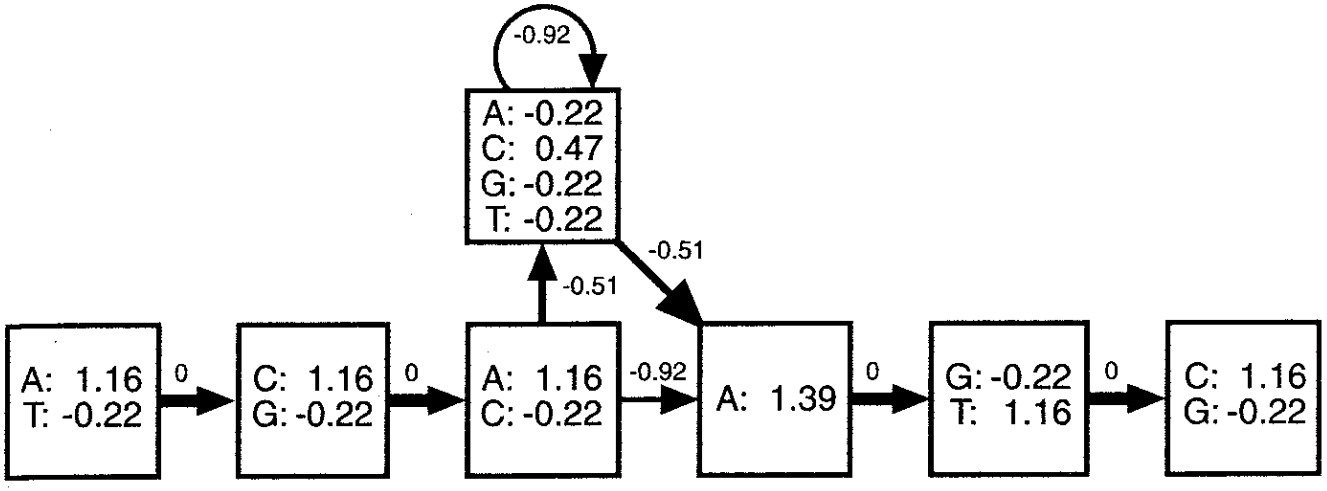
$$S_S = \ln(4p_S)$$



Построение множественного выравнивания

```
ACA---ATC
TCAACTATC
ACAC--AGC
AGA---ATC
ACCG--ATG
CGCGT-CGG
```

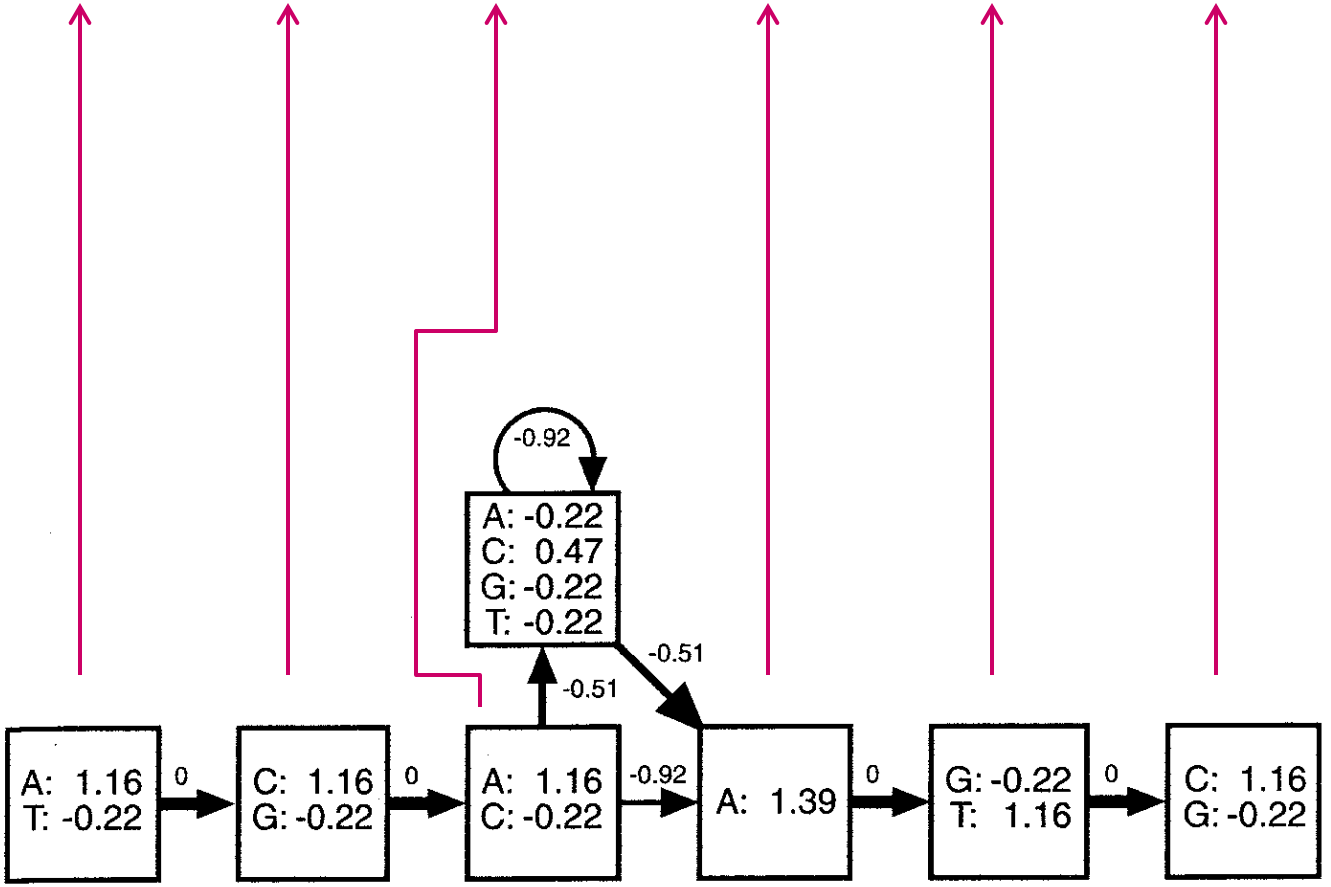
Посчитаем: описывает ли построенная модель новую последовательность?



Построение множественного выравнивания

ACA---ATC
TCAACTATC **?**
ACAC--AGC
AGA---ATC
ACCG--ATG
CGCGT-CGG **?**

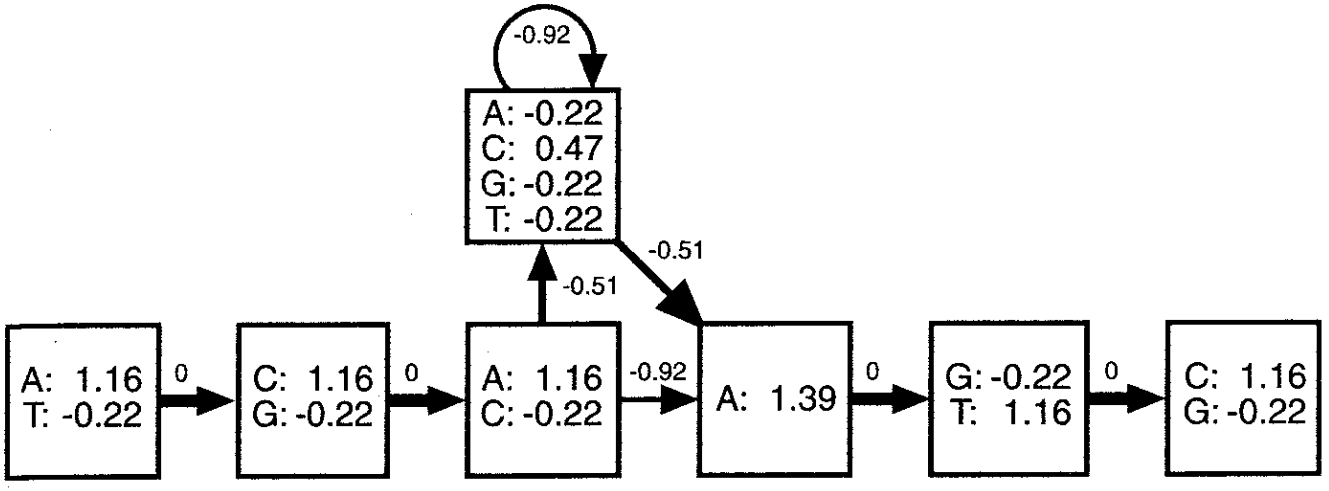
S = 1.16 + 0 + 1.16 + 0 + 1.16 - 0.92 + 1.39 + 0 + 1.16 + 0 + 1.16 = 6.29



C: 0

Построение множественного выравнивания

ACA---ATC	6.29
TCAACTATC	2.99
ACAC--AGC	5.26
AGA---ATC	4.90
ACCG--ATG	3.18
CGCGT-CGG	-3.28

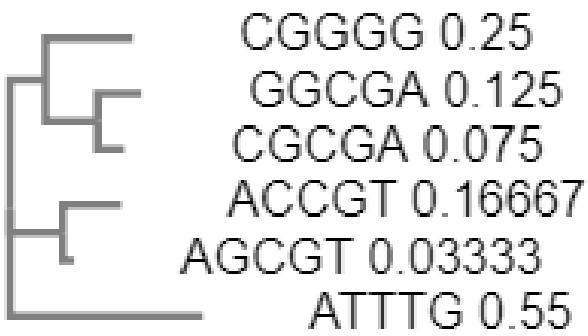


Clustal Omega

```
CLUSTAL O(1.2.1) multiple sequence alignment
```

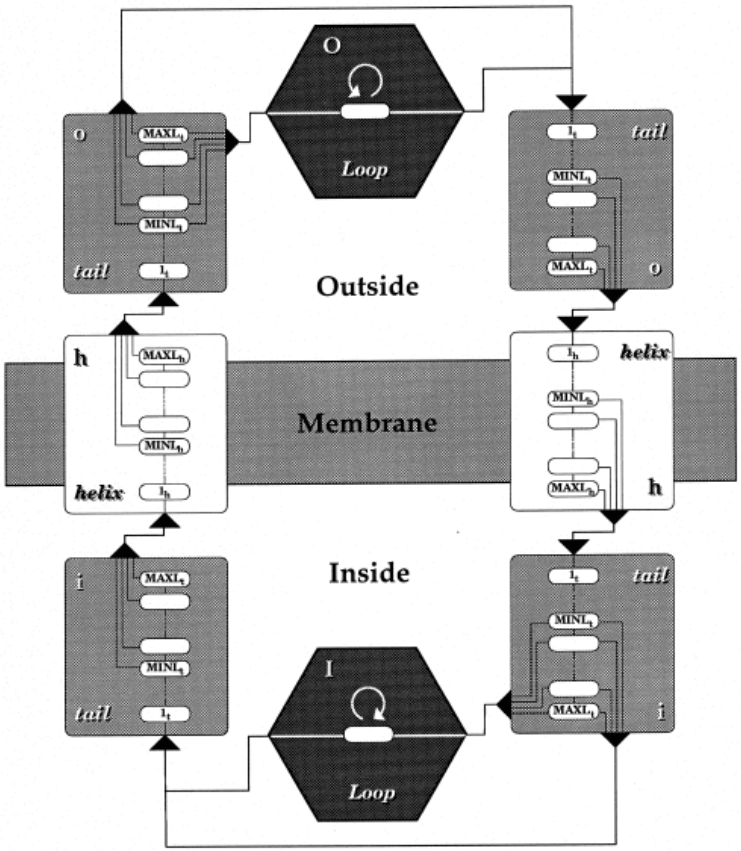
http://www.ebi.ac.uk/Tools/msa/clustalo/


```
CGGGG      CGGGG
GGCGA      GGCGA
CGCGA      CGCGA
ACCGT      ACCGT
AGCGT      AGCGT
ATTTG      ATTTG
```

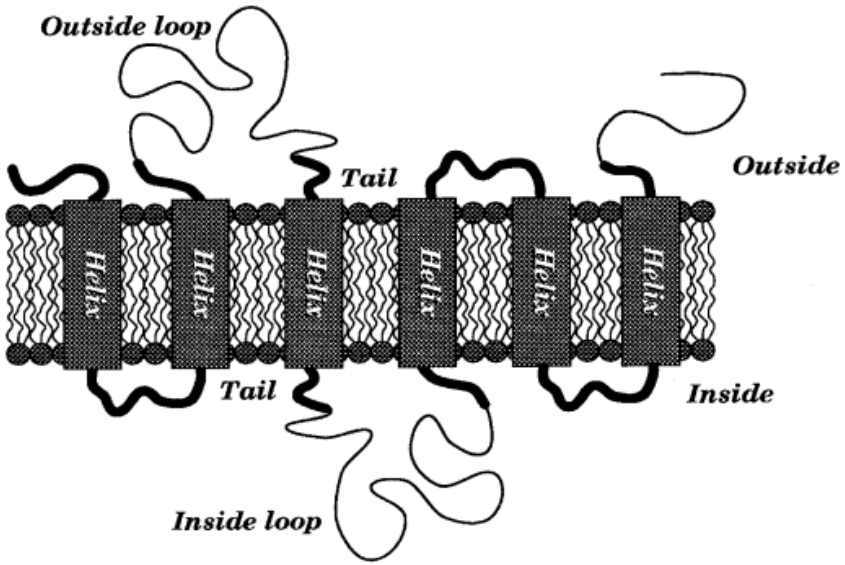


This is a Neighbour-joining tree without distance corrections.

Предсказание трансмембранных сегментов



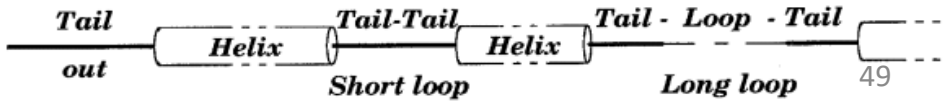
HMMTOP (1998)



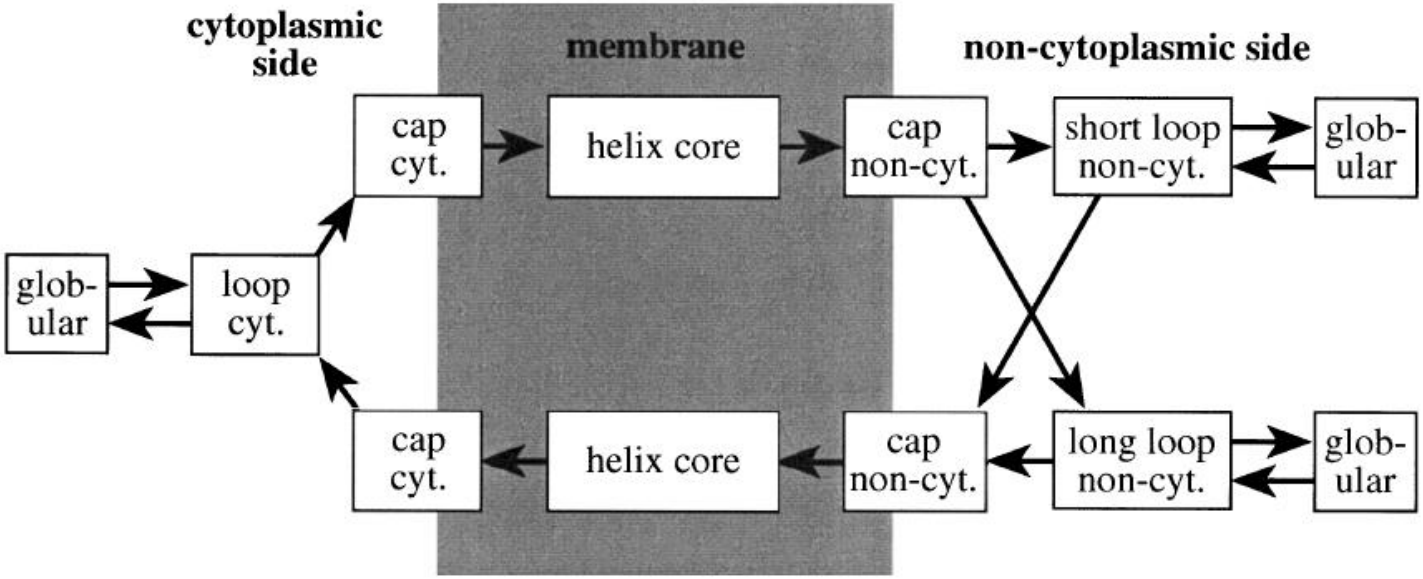
Amino acid seq: MGDVCDTEFGILVA...SVALRPRKHGRWIV...FWVDNGTEQ...PEHMTKLHMM...

State seq: oooooooooohhhhh...hhhhiiiiiiihhh...hhhooooOO...OOOoooohhh...

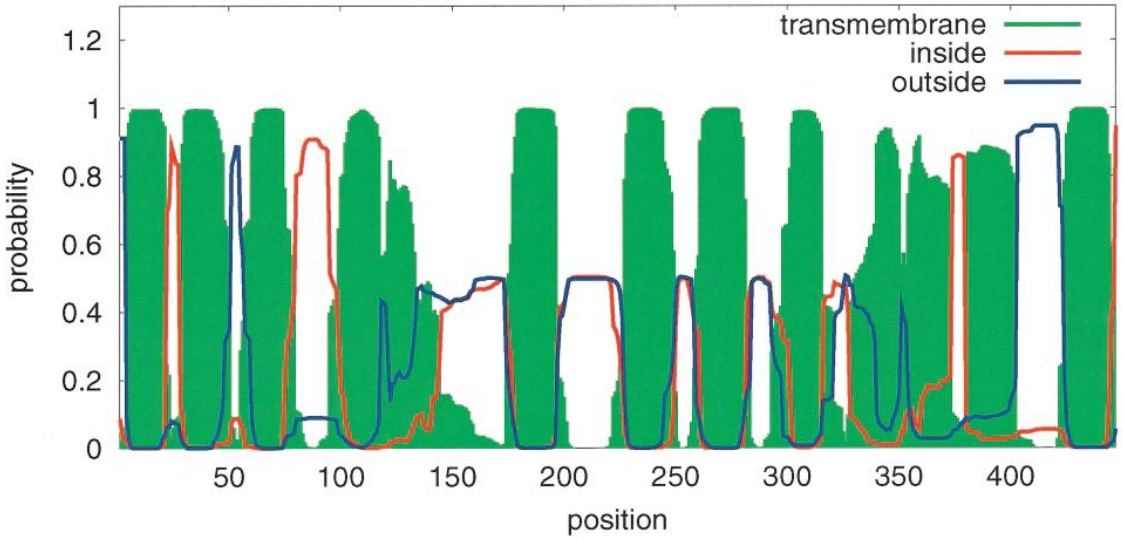
Topology:



Предсказание трансмембранных сегментов

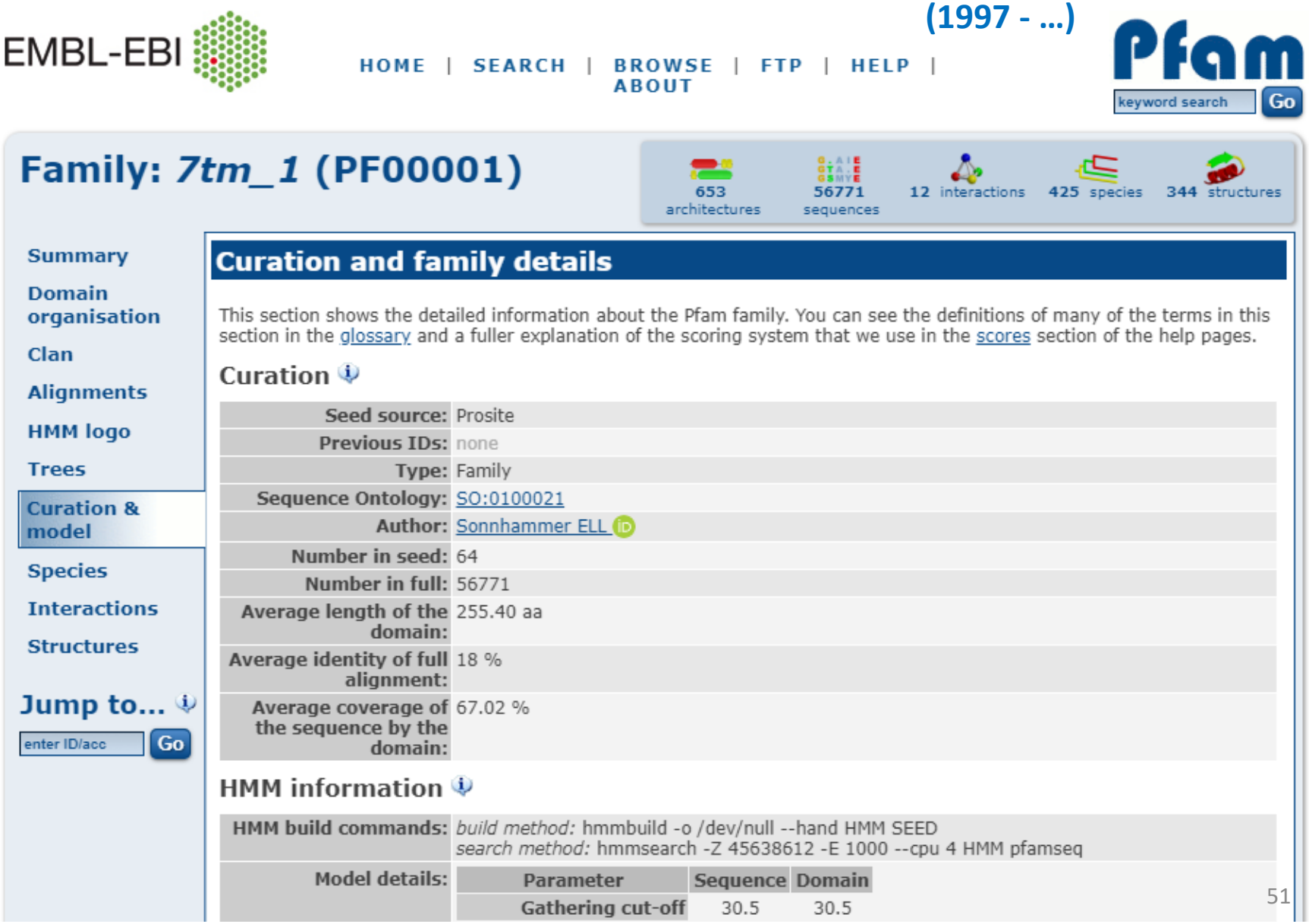


TMHMM (2001 - …)



Классификация белков по топологии

(1997 - …)



Пакет программ для анализа белковых и нуклеотидных последовательностей

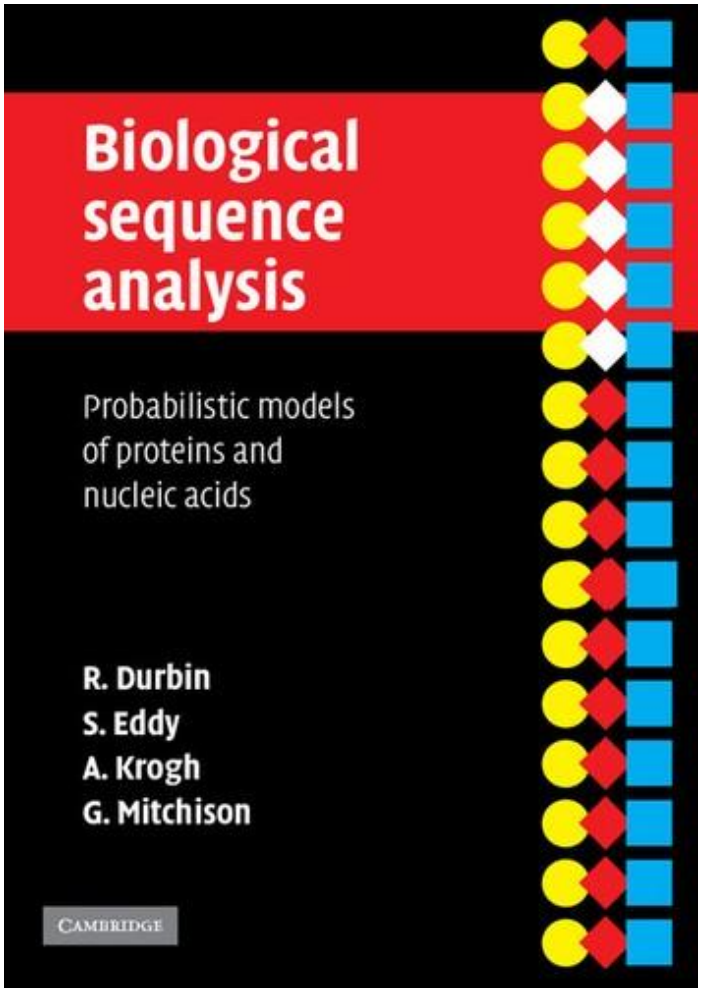
DOWNLOAD DOCUMENTATION SEARCH PUBLICATIONS BLOG

HMMER: biosequence analysis using profile hidden Markov models

Get the latest version

v3.2.1

Что почитать?



http://bioresearch.byu.edu/bio465/Labs/hmm.shtml

Множественное выравнивание последовательностей

Цели:

- **Построение филогенетических деревьев**
- **Выявление консервативных остатков и мотивов**
- **Построение профилей (визуализация)**
- **Итеративное выявление удаленной гомологии**
- …

Алгоритмы:

- Динамическое программирование – не годится
- **Прогрессивное выравнивание**
- **Скрытые марковские модели**
- Квантовые компьютеры?

Благодарю за внимание!